Apprentissage Statistique

Détection de structures communautaires dans des réseaux

Rédigé par

PRALON Nicolas

CÔME Olivier

SENE Assane

Table des matières

Introduction

De multiple réseaux, y compris les réseaux sociaux, les réseaux informatiques, se divisent plus ou moins naturellement en communautés. La détection de cette structure sous-jacente aux réseaux constitue un problème actuel, et de nombreuses approches ont été développées pour y répondre.

Dans ce rapport nous allons présenter une approche communément utilisée en apprentissage non supervisé, permettant de quantifier de la validité d'un partitionnement du réseau, les déffaillances à cette approche et la mise en pratique des méthodes utilisées pour y répondre.

Concept de Modularité

L'étude d'éventuelles structures communautaires dans des réseaux peut formellement être présentée par l'étude de graphe. Ainsi nous considérons un réseau comme un graphe, et émmettons certaines hypothèses à notre étude :

$$Soit\ G = (V, E), \text{ un graphe tels que}$$

$V = \{v_1, \cdots, v_p\}$ l'ensemble des noeuds
$E \subset \{(v_i, v_j)_{i,j \in \{1,\cdots,p\}} | i \neq j\} = V \times V$ l'ensemble des arêtes du graphe

G est un graphe simple, non orienté, non pondéré, non labélisé.

Avant de décrire l'idée mise en oeuvre pour la détection de communautés, donnons quelques définissons.

Définition 1 (Densité)**.** On appel densité d'un graphe la valeur

$$D_G = \frac{|E|}{\frac{p^2-p}{2}}$$

La densité d'un graphe correspond à la fréquence d'arêtes dans le graphe, il rend compte de la connexion entre les noeuds.

Définition 2 (Degré)**.** On appel degré d'un noeud i la valeur

$$d_i = |\{(v_i, v_j) \in E | j \in \{1, \cdots, p\}\}|$$

et correspond au nombre de voisin du noeud i.

Définition 3 (Model nul)**.** On appel model nul d'un graph G, le graph G^* dont les $|E| = m$ arêtes ont été distribuées aléatoirement entre les noeuds de G

Le model nul joue un model de référence pour lequel il n'existe aucune structure communautaire dans le réseau.

Revenons sur la question de détection de communautés et abordons là par étape.
Soit G un graphe.
Par simplicité nous souhaitons déterminer si il existe une division du graphe G en deux communautés. Une approche intuitive est de chercher deux groupes de noeuds pour lesquel on cherche à maximiser le nombre d'arête entre les neouds du groupes et à minimiser le nombre d'arête entre noeuds de groupe différent.

Cependant le simple comptage des arêtes n'est pas un bon moyen de quantifier le concept de structure communautaire. Si l'on chosissait pour groupe le graphe et pour second groupe un ensemble vide, ce partionnement seraît alors optimal, mais ne répond pas au problème.

Le concept de modularité que nous allons présenter est l'idée sous laquelle, si il existait eventuellement une structure dans un graphe, alors en le comparant à son model nul pour lequel il n'existe aucune structure sous-jacente, nous devrions raisonnablement observer une différence entre les deux models.

Définition 4 (Modularité)**.** Soit $(C_1, \cdots, C_K)$ une partition de V
On définit la modularité Q de la partition comme

$$Q(C_1, \cdots, C_K) = \frac{1}{2m} \sum_{k=1}^{K} \sum_{(v_i, v_j) \in C_k} (\mathbb{1}_{v_i, v_j) \in E} - P_{i,j})$$

avec $P_{i,j} = \frac{d_i d_j}{2m}$ la probabilité que i et j soient connectés sous le model nul, et $2m = \sum_{i=1}^{p} d_i$

La partition $(C_1, \cdots, C_K)$ est une bonne partition, selon la valeur de sa modularité, si la densité obervée dans chacun des groupes est plus élevé que la densité attendue de ce groupe, dans le model nul. Il s'agit alors de maximiser la modularité pour déterminer du meilleur partionnement, selon cette idée.